Bayesian neural networks increasingly sparsify their units with depth

Mariia Vladimirova, Julyan Arbel, Pablo Mesejo

Bayesian neural networks increasingly sparsify their units with depth

Mariia Vladimirova
Inria Grenoble Rhône-Alpes[1]
mariia.vladimirova@inria.fr

Julyan Arbel
Inria Grenoble Rhône-Alpes[1]
julyan.arbel@inria.fr

Pablo Mesejo
University of Granada
pmesejo@decsai.ugr.es

Abstract

We investigate deep Bayesian neural networks with Gaussian priors on the weights and ReLU-like nonlinearities, shedding light on novel sparsity-inducing mechanisms at the level of the units of the network, both pre- and post-nonlinearities. The main thrust of the paper is to establish that the units prior distribution becomes increasingly heavy-tailed with depth. We show that first layer units are Gaussian, second layer units are sub-Exponential, and we introduce sub-Weibull distributions to characterize the deeper layers units. Bayesian neural networks with Gaussian priors are well known to induce the *weight decay* penalty on the *weights*. In contrast, our result indicates a more elaborate regularization scheme at the level of the *units*, ranging from convex penalties for the first two layers — weight decay for the first and Lasso for the second — to non convex penalties for deeper layers. Thus, despite weight decay does not allow for the *weights* to be set exactly to zero, sparse solutions tend to be selected for the *units* from the second layer onward. This result provides new theoretical insight on deep Bayesian neural networks, underpinning their natural shrinkage properties and practical potential.

[1]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

1 Introduction

Neural networks (NNs) (Bishop, 1995), and their deep extensions (Goodfellow et al., 2016), have largely been used in many research areas such as image analysis (Krizhevsky et al., 2012), signal processing (Graves et al., 2013), or reinforcement learning (Silver et al., 2016), just to name a few. These performances have greatly strengthened the line of research that aims at better understanding the driving mechanisms behind the effectiveness of deep neural networks. One important aspect of this analysis that has recently gained much attention is the study of distributional properties of the NNs through Bayesian inference.

Bayesian approaches investigate models by assuming a prior distribution on their parameters. Bayesian machine learning refers to extending standard machine learning approaches with posterior inference, a line of research pioneered by the works Neal (1992); MacKay (1992) on Bayesian neural networks which now extends to a broad class of models, including Bayesian GAN (Saatci and Wilson, 2017). See Polson and Sokolov (2017) for a review. The interest of the Bayesian approach to NNs is at least twofold. First, it offers a principled approach for modeling uncertainty of the training procedure, which is a limitation of standard NNs which only provide point estimates. A second main asset of Bayesian models is that they represent regularized versions of their classical counterparts. For instance, mode a posteriori (MAP) estimation of a Bayesian regression model with double exponential (Laplace) prior is equivalent to Lasso regression (Tibshirani, 1996), while a Gaussian prior leads to ridge regression. When it comes to neural networks, the regularization mechanism is also well appreciated in the literature, since neural networks traditionally suffer from overparameterization, resulting in overfitting.

Central in the field of regularization techniques is the *weight decay* penalty (Krogh and Hertz, 1991), which is equivalent to MAP estimation of a Bayesian neural network with independent Gaussian priors on the weights. Srivastava et al. (2014) have suggested *dropout* as a regularization method in which neurons are randomly turned off. Gal and Ghahramani (2016) proved that the neural network trained with *dropout* is equivalent to a probabilistic model, i.e. a deep Gaussian process (Damianou and Lawrence, 2013). It leads to the consideration of such neural networks as Bayesian models.

This study is devoted to the investigation of hidden units prior distributions in Bayesian neural networks under assumption of independent Gaussian weights. We first describe a fully connected neural network architecture as illustrated in Figure 1. Given an input $\boldsymbol{x} \in \mathbb{R}^N$, the ℓ-th hidden layer unit activations are defined as

$$\boldsymbol{g}^{(\ell)}(\boldsymbol{x}) = \boldsymbol{W}^{(\ell)}\boldsymbol{h}^{(\ell-1)}(\boldsymbol{x}), \quad \boldsymbol{h}^{(\ell)}(\boldsymbol{x}) = \phi(\boldsymbol{g}^{(\ell)}(\boldsymbol{x})), \tag{1}$$

where $\boldsymbol{W}^{(\ell)}$ is a weight matrix including the bias vector. A nonlinear activation function $\phi : \mathbb{R} \to \mathbb{R}$ is applied element-wise, which is called nonlinearity, $\boldsymbol{g}^{(\ell)} = \boldsymbol{g}^{(\ell)}(\boldsymbol{x})$ is a vector of pre-nonlinearities, and $\boldsymbol{h}^{(\ell)} = \boldsymbol{h}^{(\ell)}(\boldsymbol{x})$ is a vector of post-nonlinearities. When we refer to either pre- or post-nonlinearities, we will use the notation $\boldsymbol{U}^{(\ell)}$.

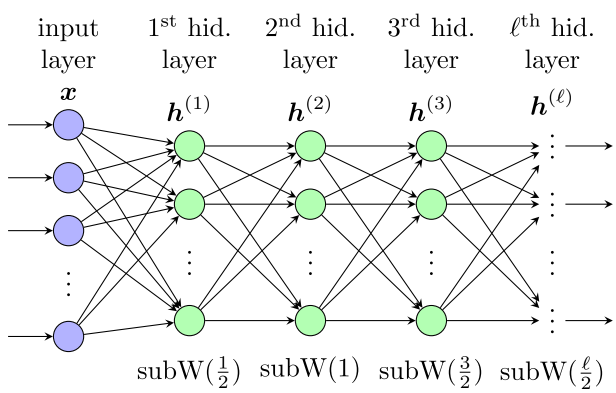

Figure 1: Neural network architecture and characterization of the ℓ-layer units prior distribution as sub-Weibull distribution with tail parameter $\ell/2$ (see Definition 4.1).

1.1 Contributions

In this paper, we extend the theoretical understanding of feedforward fully connected neural networks by studying prior distributions at the units level, under the assumption of independent and normally distributed weights. Our contributions are threefold:

(i) We define the notion of *sub-Weibull* distributions (Definition 4.1), which are characterized by tails lighter than (or equally light as) Weibull distributions; in the same way as sub-Gaussian or sub-Exponential distributions correspond to distributions with tails lighter than Gaussian and Exponential distributions, respectively. Sub-Weibull distributions are parameterized by a positive tail index θ and equivalent to sub-Gaussian for $\theta = 1/2$ and sub-Exponential for $\theta = 1$. We provide a moments characterization of the sub-Weibull property in Proposition 4.1.

(ii) As our main contribution, we prove in Theorem 3.1 that under some conditions on the nonlinear function ϕ, a Gaussian prior on the weights induces a sub-Weibull distribution on the units (both pre- and post-nonlinearities) with optimal tail parameter $\theta = \ell/2$ (see Figure 1). The condition on ϕ essentially imposes that ϕ strikes to $+\infty$ or $-\infty$ for large absolute values of the argument, as ReLU does. In the case of bounded support ϕ, like *sigmoid* or *tanh*, the units are bounded, making them *de facto* sub-Gaussian[2].

(iii) Lastly, we offer an interpretation of the main result from a sparsity-inducing viewpoint in Section 2. Heavy-tailed priors are known to induce a sparse model representation, such as the Lasso in a regression problem. We extrapolate this finding to the setting of a deep Bayesian neural network (BNN), showing that the units tend to be more sparsely represented as layers become deeper.

1.2 Related work

Studying the distributional behaviour of feedforward networks has been a fruitful avenue for understanding these models, as pioneered by the works of Radford Neal (Neal, 1992, 1996) and David MacKay (MacKay, 1992). The first results in the field addressed the limiting setting when the number of units per layer tends to infinity, also called the wide regime. Neal (1996) proved that a single hidden layer neural network with normally distributed weights tends in distribution in the wide limit either to a Gaussian process (Rasmussen and Williams, 2006) or to an α-stable process, depending on how the prior variance on the weights is rescaled. In recent works, Matthews et al. (2018a), or its extended version Matthews et al. (2018b), and Lee et al. (2018) extend the result of Neal to more-than-one-layer neural networks: when the number of hidden units grows to infinity, deep neural networks (DNNs) also tend in distribution to the Gaussian process, under the assumption of Gaussian weights for

[2] A trivial version of our main result holds.

properly rescaled prior variances. For the rectified linear unit (ReLU) activation function, the Gaussian process covariance function is obtained analytically (Cho and Saul, 2009). For other nonlinear activation functions, Lee et al. (2018) use a numerical approximation algorithm.

Various distributional properties are also studied in neural networks regularization methods. The *dropout* technique (Srivastava et al., 2014) was reinterpreted as a form of approximate Bayesian variational inference (Kingma et al., 2015; Gal and Ghahramani, 2016). Gal and Ghahramani (2016) built a connection between dropout and the Gaussian process. Kingma et al. (2015) proposed a way to interpret Gaussian dropout. They suggested *variational dropout* where each weight of a model has its individual dropout rate. *Sparse variational dropout* from Molchanov et al. (2017) extends *variational dropout* to all possible values of dropout rates and leads to a sparse solution. The approximate posterior is chosen to factorize either over rows or individual entries of the weight matrices. The prior usually factorizes in the same way. The choice of the prior and its interaction with the approximating posterior family are studied in Hron et al. (2018). Performing dropout can be used as a Bayesian approximation. However, as noted by Duvenaud et al. (2014), dropout has no regularization effect on infinitely-wide hidden layers.

The recent work by Bibi et al. (2018) provides the expression of the first two moments of the output units of a one layer neural network. Obtaining the moments is a first step to characterizing a whole distribution, however the methodology of Bibi et al. (2018) is limited to the first two moments and to one layer neural networks, while we address the problem in more generality for deep neural networks.

In the remainder of the paper, we present our three main contributions in reverse but perhaps more pedagogical order, that is starting with intuitions and interpretation (iii), then moving to theoretical results (ii) and ending up with the necessary statistical background and definitions (i). More specifically, Section 2 illustrates shrinkage and penalization techniques, providing an interpretation for our main contribution, Theorem 3.1, that is stated in Section 3. Section 4 introduces the sub-Weibull distribution family based on its tail behavior and proves a moment-based characterization. Conclusions and future work are reported in Section 5, while proofs and additional technical results are deferred to the Supplementary material.

2 Sparsity-inducing prior on the units

2.1 Short digest on penalized estimation

Our main theoretical contribution, Theorem 3.1, characterizes the marginal prior distribution of the network units as follows: when the depth increases, the distribution becomes more heavy-tailed, as will be precised in the next section. First, in this section, we provide an interpretation of the result in terms of sparsity-inducing mechanism at the level of the units. To this aim, a short reminder about shrinkage methods and penalization is presented.

The shrinkage idea is probably best illustrated on the simple linear regression model, where the aim of shrinkage is to improve prediction accuracy by shrinking, or even putting exactly to zero, some coefficients in the regression. Under these circumstances, inference is also more *interpretable* since, by reducing the number of coefficients effectively used in the model, it is possible to grasp its salient features. Shrinking is performed by imposing a penalty on the size of the coefficients, which is equivalent to allowing for a given budget on their size. Denote the regression parameter by $\beta \in \mathbb{R}^p$, the regression sum-of-squares by $R(\beta)$, and the penalty by $\lambda L(\beta)$, where L is some norm on $\mathbb{R}^p$ and λ some positive tuning parameter. Then, the two formulations of the regularized problem

$$\min_{\beta\in\mathbb{R}^p} R(\beta) + \lambda L(\beta), \text{ and}$$
$$\min_{\beta\in\mathbb{R}^p} R(\beta) \text{ subject to } L(\beta) \le t,$$

are equivalent, with some one-to-one correspondence between λ and t, and are respectively termed the *penalty* and the *constraint* formulation. This latter formulation provides an interesting geometrical intuition of the shrinkage mechanism: the constraint $L(\beta) \le t$ reads as imposing a total budget of t for the parameter size in terms of the norm L. If the ordinary least squares estimator $\hat{\beta}^{\text{ols}}$ lives in the L-ball with surface $L(\beta) = t$, then there is no effect on the estimation. In contrast, when $\hat{\beta}^{\text{ols}}$ is outside the ball, then the intersection of the lowest level curve of the sum-of-squares $R(\beta)$ with the L-ball defines the penalized estimator.

The choice of the L norm has considerable effects on the problem, as can be sensed geometrically. Consider for instance L^q norms, with $q \ge 0$. For any $q > 1$, the associated L^q norm is differentiable and contours have a round shape without sharp angles. In that case, the penalty effect is to shrink the β coefficients towards 0. The most well-known estimator falling in this class is the *ridge* regression obtained with $q = 2$, see Figure 2 top-left panel. In contrast, for any $q \in (0, 1]$, the

L^q norm has some non differentiable points along the axis coordinates, see Figure 2 top-right and bottom panels. Such critical points are more likely to be hit by the level curves of the sum-of-squares $R(\beta)$, thus setting exactly to zero some of the parameters. A very successful approach in this class is the Lasso obtained with $q = 1$. Note that the problem is computationally much easier in the convex situation which occurs only for $q \geq 1$.

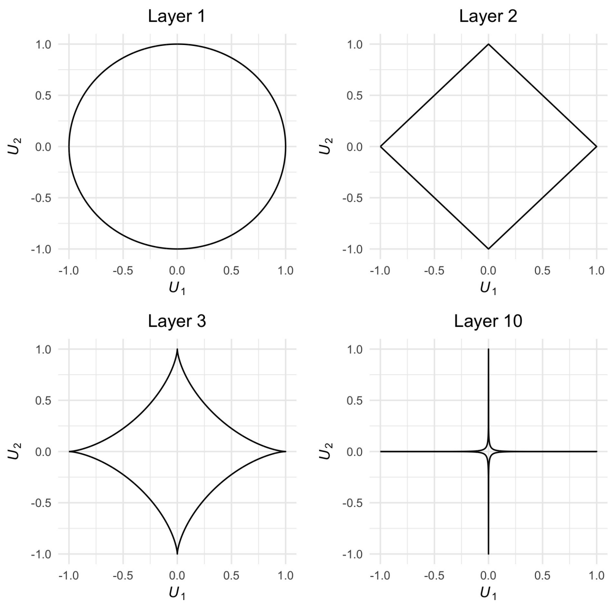

Figure 2: $\mathcal{L}^{2/\ell}$-norm unit balls (in dimension 2) for layers $\ell = 1, 2, 3$ and 10.

2.2 MAP on weights $\boldsymbol{W}$ is weight decay

These penalized methods have a simple Bayesian counterpart in the form of the mode a posteriori (MAP) estimator. In this context, the objective function R is the negative log-likelihood, while the penalty L is the negative log-prior. The objective function takes on the form of sum-of-squared errors for regression under Gaussian errors, and of cross-entropy for classification.

For neural networks, it is well-known that an independent Gaussian prior on the weights

$$\pi(\boldsymbol{W}) = \prod_{\ell=1}^{L} \prod_{i,j} \mathrm{e}^{-\frac{1}{2}(W_{i,j}^{(\ell)})^2},$$

is equivalent to the weight decay penalty, also known as ridge regression:

$$L(\boldsymbol{W}) = \sum_{\ell=1}^{L} \sum_{i,j} (W_{i,j}^{(\ell)})^2 = \|\boldsymbol{W}\|_2^2,$$

where products and sums involving i and j above are over $1 \leq i \leq H_{\ell-1}$ and $1 \leq j \leq H_\ell$, H_0 and H_L representing respectively the input and output dimensions.

2.3 MAP on units $\boldsymbol{U}$ induces sparsity

Now moving the point of view from *weights* to *units* leads to a radically different shrinkage effect. Let $U_m^{(\ell)}$ denote the m-th unit of the ℓ-th layer (either pre- or post-nonlinearity). We prove in Theorem 3.1 that conditional on the input $\boldsymbol{x}$, a Gaussian prior on the weights translates into some prior on the units $U_m^{(\ell)}$ that is marginally sub-Weibull with optimal tail index $\theta = \ell/2$. This means that the tails of $U_m^{(\ell)}$ satisfy

$$\mathbb{P}(|U_m^{(\ell)}| \geq u) \leq \exp\left(-u^{2/\ell}/K_1\right) \quad \text{for all } u \geq 0, \quad (2)$$

for some positive constant K_1. The exponent of u in the exponential term above is optimal in the sense that Equation (2) is not satisfied with some parameter θ' smaller than $\ell/2$. Thus, the marginal density of $U_m^{(\ell)}$ on $\mathbb{R}$ is approximately proportional to

$$\pi_m^{(\ell)}(u) \approx \mathrm{e}^{-u^{2/\ell}/K_1}.$$

The joint prior distribution for all the units $\boldsymbol{U} = (U_m^{(\ell)})_{1\leq\ell\leq L, 1\leq m\leq H_\ell}$ can be expressed from all the marginal distributions by Sklar's representation theorem as

$$\pi(\boldsymbol{U}) = \prod_{\ell=1}^{L} \prod_{m=1}^{H_\ell} \pi_m^{(\ell)}(U_m^{(\ell)})\, C(F(\boldsymbol{U})), \quad (3)$$

where C represents the copula of $\boldsymbol{U}$ (which characterizes all the dependence between the units) while F denotes its cumulative distribution function. The penalty incurred by such a prior distribution is obtained as the negative log-prior,

$$\begin{aligned} L(\boldsymbol{U}) &= -\sum_{\ell=1}^{L} \sum_{m=1}^{H_\ell} \log \pi_m^{(\ell)}(U_m^{(\ell)}) - \log C(F(\boldsymbol{U})), \\ &\approx \sum_{\ell=1}^{L} \sum_{m=1}^{H_\ell} |U_m^{(\ell)}|^{2/\ell} - \log C(F(\boldsymbol{U})), \\ &\approx \|\boldsymbol{U}^{(1)}\|_2^2 + \|\boldsymbol{U}_1^{(2)}\|_1 + \cdots + \|\boldsymbol{U}^{(L)}\|_{2/L}^{2/L} \\ &\quad - \log C(F(\boldsymbol{U})). \end{aligned} \quad (4)$$

The first L terms in (4) indicate that some shrinkage operates at every layer of the network, with a penalty term that takes the form of the $\mathcal{L}^{2/\ell}$ norm. Thus, the deeper the layer, the stronger the sparsity at the level of the units, as summarized in Table 1.

Layer	Penalty on $\boldsymbol{W}$	Penalty on $\boldsymbol{U}$	
1	$\|\boldsymbol{W}^{(1)}\|_2^2$, $\mathcal{L}^2$	$\|\boldsymbol{U}^{(1)}\|_2^2$	$\mathcal{L}^2$ (weight decay)
2	$\|\boldsymbol{W}^{(2)}\|_2^2$, $\mathcal{L}^2$	$\|\boldsymbol{U}^{(2)}\|$	$\mathcal{L}^1$ (Lasso)
ℓ	$\|\boldsymbol{W}^{(\ell)}\|_2^2$, $\mathcal{L}^2$	$\|\boldsymbol{U}^{(\ell)}\|_{2/\ell}^{2/\ell}$	$\mathcal{L}^{2/\ell}$

Table 1: Comparison of Bayesian neural network shrinkage effect on weights $\boldsymbol{W}$ and units $\boldsymbol{U}$.

3 Bayesian neural networks have heavy-tailed deep units

The deep learning approach uses stochastic gradient descent and error back-propagation in order to fit the network parameters $(\mathbf{W}^{(\ell)})_{1\leq\ell\leq L}$, where ℓ iterates over all network layers. In the Bayesian approach, the parameters are random variables described by probability distributions.

3.1 Assumptions on neural network

We assume a prior distribution on the model parameters, that are the weights $\boldsymbol{W}$. In particular, let all weights (including biases) be independent and have zero-mean normal distribution

$$W_{i,j}^{(\ell)} \sim \mathcal{N}(0, \sigma_w^2), \tag{5}$$

for all $1 \leq \ell \leq L$, $1 \leq i \leq H_{\ell-1}$ and $1 \leq j \leq H_\ell$. Given some input $\boldsymbol{x}$, such prior distribution induces by forward propagation 1 a prior distribution on the pre-nonlinearities and post-nonlinearities, whose *tail properties* are the focus of this section. To this aim, the nonlinearity ϕ is required to span at least half of the real line as follows. We introduce an extended version of the nonlinearity assumption from Matthews et al. (2018b):

Definition 3.1 (Extended envelope property for nonlinearities)**.** *A nonlinearity $\phi : \mathbb{R} \to \mathbb{R}$ is said to obey the extended envelope property if there exist $c_1, c_2, d_1, d_2 \geq 0$ such that the following inequalities hold*

$$\begin{aligned} |\phi(u)| &\geq c_1 + d_1|u| \quad \textit{for all } u \in \mathbb{R}_+ \textit{ or } u \in \mathbb{R}_-, \\ |\phi(u)| &\leq c_2 + d_2|u| \quad \textit{for all } u \in \mathbb{R}. \end{aligned} \tag{6}$$

The interpretation of this property is that ϕ must shoot to infinity at least in one direction ($\mathbb{R}_+$ or $\mathbb{R}_-$, at least linearly (first line of (6)), and also at most linearly (second line of (6)). Of course, compactly supported nonlinearities such as sigmoid and tanh do not satisfy the extended envelope property but the majority of other nonlinearities do, including ReLU, ELU, PReLU, and SeLU.

Lemma 3.1. *Let a nonlinearity $\phi : \mathbb{R} \to \mathbb{R}$ obey the extended envelope property. Then for any symmetric random variable X the following asymptotic equivalence*[3] *holds*

$$\mathbb{E}[\phi(X)^k] \asymp \mathbb{E}[X^k], \quad \textit{for } k \to \infty. \tag{7}$$

The proof is deferred to the Supplementary material.

3.2 Main theorem

This section postulates the rigorous result with a proof sketch. In Supplementary material one can find proofs of intermediate lemmas and a covariance theorem which states the non-negative covariance between post-nonlinearities.

Theorem 3.1 (Sub-Weibull units)**.** *Consider a feed-forward Bayesian neural network with Gaussian priors (5) with nonlinearity ϕ satisfying the extended envelope condition of Definition 3.1. Then conditional on the input $\boldsymbol{x}$, the marginal prior distribution induced by forward propagation (1) on any unit (pre- or post-nonlinearity) of the ℓ-th hidden layer is sub-Weibull with optimal tail parameter $\theta = \ell/2$. That is for any $1 \leq \ell \leq L$, and for any $1 \leq m \leq H_\ell$,*

$$U_m^{(\ell)} \sim \mathsf{subW}(\ell/2),$$

where a subW *distribution is defined in Definition 4.1, and $U_m^{(\ell)}$ is either a pre-nonlinearity $g_m^{(\ell)}$ or a post-nonlinearity $h_m^{(\ell)}$.*

Proof. The idea is to prove by induction with respect to hidden layer depth ℓ that pre- and post-nonlinearities satisfy the asymptotic moment equivalence

$$\|g^{(\ell)}\|_k \asymp k^{\ell/2} \text{ and } \|h^{(\ell)}\|_k \asymp k^{\ell/2}.$$

The statement of the theorem then follows by the moment characterization of optimal sub-Weibull tail coefficient in Proposition 4.3.

According to Lemma 4.1, centering does not harm tails properties, then, for simplicity, we consider zero-mean distributions $W_{i,j}^{(\ell)} \sim \mathcal{N}(0, \sigma_w^2)$.

Base step: consider the distribution of the first hidden layer pre-nonlinearity g ($\ell = 1$). Since weights $\boldsymbol{W}_m$ follow normal distribution and $\boldsymbol{x}$ is a feature vector, then each hidden unit $\boldsymbol{W}_m^\top \boldsymbol{x}$ follow also normal distribution

$$g = \boldsymbol{W}_m^\top \boldsymbol{x} \sim \mathcal{N}(0, \sigma_w^2 \|\boldsymbol{x}\|^2).$$

Then, for normal zero-mean variable g, having variance $\sigma^2 = \sigma_w^2\|\boldsymbol{x}\|^2$, it holds the equality in sub-

Gaussian property with variance proxy equals to normal distribution variance and from Lemma B.1:

$$\|g\|_k \asymp \sqrt{k}.$$

As activation function ϕ obeys extended envelope property, according to Lemma B.2, nonlinearity moments are asymptotic equivalent to symmetric variable moments

$$\|\phi(g)\|_k \asymp \|g\|_k \sim \sqrt{k}.$$

It implies that first hidden layer post-nonlinearities h have sub-Gaussian distribution or sub-Weibull with tail parameter $\theta = 1/2$ (Definition 4.1).

Inductive step: show that if the statement holds for $\ell - 1$, then it also holds for ℓ.

Suppose the post-nonlinearity of $(\ell-1)$-th hidden layer satisfies the moment condition. Hidden units satisfy the non-negative covariance theorem (Theorem C.1):

$$\operatorname{Cov}\left[\left(h^{(\ell-1)}\right)^s, \left(\tilde{h}^{(\ell-1)}\right)^t\right] \geq 0, \text{ for any } s, t \in \mathbb{N}.$$

Let the number of hidden units in $(\ell-1)$-th layer equals to H. Then according to Lemma B.3, under assumption of zero-mean Gaussian weights, pre-nonlinearities of ℓ-th hidden layer $g^{(\ell)} = \sum_{i=1}^H W_{m,i}^{(\ell-1)} h_i^{(\ell-1)}$ also satisfy the moment condition, but with $\theta = \ell/2$

$$\|g^{(\ell)}\|_k \asymp k^{\ell/2}.$$

Using Lemma B.2 from Supplementary material, one can show that post-nonlinearities $h^{(\ell)}$ satisfy the same moment condition as pre-nonlinearities $g^{(\ell)}$. This finishes the proof. □

We illustrate the result of Theorem 3.1 in Figure 3 which represents the first three hidden layers pre-nonlinearity marginal distributions (top panel). These densities are obtained as kernel density estimators from a sample of size 10^5 from the prior on the pre-nonlinearities, which is itself obtained by sampling 10^5 sets of weights $\boldsymbol{W}$ from the Gaussian prior (5) and forward propagation via (1). The three hidden layers of neural network have $H_1 = 25$, $H_2 = 24$ and $H_3 = 4$ hidden units, respectively. Being a linear combination involving symmetric weights $\boldsymbol{W}$, pre-nonlinearities $\boldsymbol{g}$ are also symmetric, thus we visualize only their positive part. The input vector $\boldsymbol{x} \in \mathbb{R}^{50}$ is sampled from a standard normal distribution once for all at the start. The nonlinearity ϕ is the ReLU function. The prior distribution of post-nonlinearities has a Dirac mass at zero with a coefficient of 1/2 and they are no more symmetric. But the post-nonlinearity prior distribution tails remain the same as of pre-nonlinearities on $\mathbb{R}_+$, and is represented on the bottom panel of Figure 3.

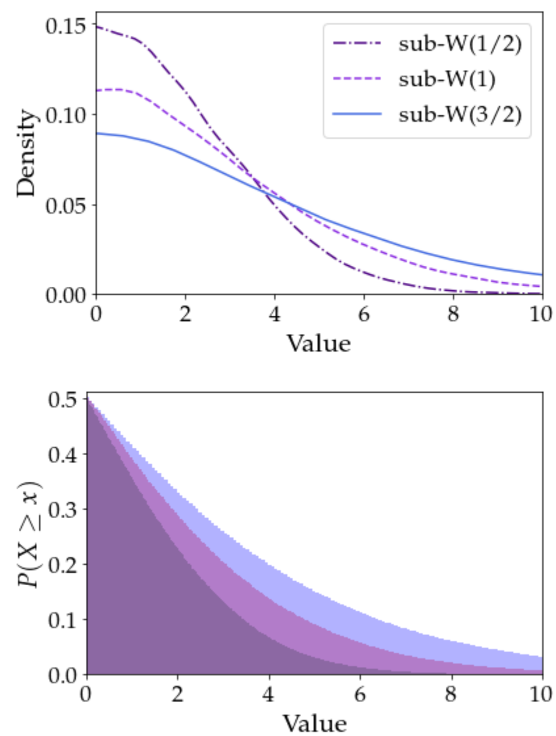

Figure 3: Illustration of the first three layers hidden units marginal prior distributions.

Remark 3.1. If the activation function ϕ is bounded, such as the sigmoid, or tanh, then the units are bounded. As a result, by Hoeffding's Lemma, they have a sub-Gaussian distribution.

3.3 Convolutional neural networks

Convolutional neural networks (Goodfellow et al., 2016) are a particular kind of neural network for processing data that has a known grid-like topology, which allows to encode certain properties into the architecture. These then make the forward function more efficient to implement and vastly reduce the amount of parameters in the neural network. Neurons in such networks are arranged in three dimensions: width, height and depth. There are three main types of layers that can be concatenated in CNN architectures: convolutional layer, pooling layer, and fully-connected layer (exactly as seen in standard Neural Networks). Convolutional layer computes dot products between a region in the inputs and its weights. Therefore, each region can be considered as a particular case of fully-connected layer. Pooling layer is performed to control overfitting and computations in deep architectures. The pooling layer operates independently on every depth slice of the input and residues it spatially.

The most commonly functions used in pooling layers are *max pooling* and *average pooling*.

Proposition 3.1. *The operations: 1.* max pooling *and 2.* averaging *do not modify the optimal tail parameter θ of sub-Weibull family. Consequently, the result of Theorem 3.1 carries over to Convolutional neural networks.*

Proof. Let $X_i \sim \mathsf{subW}(\theta)$ for $1 \leq i \leq N$ be units from one region where pooling operation is applied. Using Definition 4.1, for all $x \geq 0$ and some constant $K > 0$ we have

$$\mathbb{P}(|X_i| \geq x) \leq \exp\left(-x^{1/\theta}/K\right) \text{ for all } i.$$

1. Max pooling operation takes the maximum element in the region. Since X_i, $1 \leq i \leq N$ are the elements in one region, we want to check if the tail of $\max_{1\leq i\leq N} X_i$ obeys sub-Weibull property with optimal tail parameter equals to θ. Using probability properties, we get

$$\begin{aligned}\mathbb{P}(\max_{1\leq i\leq N} X_i \geq x) &= \mathbb{P}(\bigcup_{1\leq i\leq N} X_i \geq x) \\ &\leq \sum_{i=1}^{N} \mathbb{P}(X_i \geq x).\end{aligned}$$

   Taking into account that the variables X_i belong to one sub-Weibull family for all $1 \leq i \leq N$, the distribution tail has the following form

$$\mathbb{P}(\max_{1\leq i\leq N} X_i \geq x) \leq 2N \exp\left(-x^{1/\theta}/K\right).$$

   The constant 2 is appeared through distribution symmetry. Then, for $K_1 = K/\log(2N)$, the proposition for max pooling layer holds.

2. Apply averaging on the unit region:

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} X_i \geq x\right) = \bigcap_{1\leq i\leq N} \mathbb{P}\big(X_i/N \geq x\big).$$

   Substituting the sub-Weibull property, we obtain the same upper bound

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} X_i \geq x\right) \leq \exp\left(-x^{1/\theta}/K\right).$$

   Summarizing and division by a constant does not influence the distribution tail, yielding the proposition result regarding the averaging operation.

□

Corollary 3.1. *Consider a convolutional neural network containing convolutional, pooling and fully-connected layers under assumptions from Section 3.1. Then a unit of ℓ-th hidden layer has sub-Weibull distribution with optimal tail parameter $\theta = \ell/2$, where ℓ is the number of convolutional and fully-connected layers.*

Proof. Proposition 3.1 implies that the pooling layer keeps the tail parameter. From discussion at the beginning of the section, the result of Theorem 3.1 is also applied to CNNs where the depth is considered as the number of convolutional and fully-connected layers. □

4 Sub-Weibull distributions

Let X be a random variable. The following proposition states different equivalent distribution properties, such as tail decay and the growth of moments. The proof of this result shows how to transform one type of information about random variables into another.

Proposition 4.1 (Equivalent properties)**.** *Let X be a random variable. Then the following properties are equivalent; the parameters $K_i > 0$ appearing in these properties differ from each other by at most an absolute constant factor.*

1. *The tails of X satisfy*

$$\mathbb{P}(|X| \geq x) \leq \exp\left(-x^{1/\theta}/K_1\right) \quad \text{for all } x \geq 0.$$

2. *The moments of X satisfy*

$$\|X\|_k = \left(\mathbb{E}[|X|^k]\right)^{1/k} \leq K_2 k^{\theta} \quad \text{for all } k \geq 1.$$

It is a short version of Proposition A.1 which states two additional equivalent properties in terms of the moment generating function of the random variable X. Proposition A.1 is deferred to and proved in Supplementary material.

Definition 4.1 (Sub-Weibull random variable)**.** *A random variable X that satisfies one of the equivalent properties of Proposition 4.1 is called a sub-Weibull random variable with the tail parameter θ, which is denoted by $X \sim \mathsf{subW}(\theta)$.*

Informally, the tails of a $\mathsf{subW}(\theta)$ distribution are dominated by (i.e. decay at least as fast as) the tails of a Weibull variable with the shape parameter equal to $1/\theta$ (Rinne, 2008). The larger tail parameter θ, the heavier the tails of the sub-Weibull distribution.

Sub-Gaussian and sub-Exponential variables, which are commonly used, are special cases of sub-Weibull random variables with tail parameter $\theta = 1/2$ and $\theta = 1$, respectively (see Table 2).

Distribution	Tail	Moments
Sub-Gaussian	$\overline{F}(x) \leq \mathrm{e}^{-\lambda x^2}$	$\|X\|_k \leq C\sqrt{k}$
Sub-Exponential	$\overline{F}(x) \leq \mathrm{e}^{-\lambda x}$	$\|X\|_k \leq Ck$
Sub-Weibull	$\overline{F}(x) \leq \mathrm{e}^{-\lambda x^{1/\theta}}$	$\|X\|_k \leq Ck^\theta$

Table 2: Sub-Gaussian, sub-Exponential and sub-Weibull distributions comparison in terms of tail $\overline{F}(x) = P(X \geq x)$ and moment condition, with λ and C some positive constants. The first two are a special case of the last with $\theta = 1/2$ and $\theta = 1$ respectively.

Proposition 4.2 (Inclusion)**.** *Let θ_1 and θ_2 such that $0 < \theta_1 < \theta_2$ be tail proxy parameters for some sub-Weibull distributed variables. Then the following inclusion holds*

$$\mathsf{subW}(\theta_1) \subset \mathsf{subW}(\theta_2).$$

Proof. For $X \sim \mathsf{subW}(\theta_1)$, it holds that $\|X\|_k \leq K_2 k^{\theta_1}$. Since $k^{\theta_1} \leq k^{\theta_2}$ for all $k \geq 1$, this yields $\|X\|_k \leq K_2 k^{\theta_2}$, which by definition implies $X \sim \mathsf{subW}(\theta_2)$. □

The following proposition is key in establishing that neural network units of layer ℓ are $\mathsf{subW}(\ell/2)$, where $\ell/2$ is optimal.

Proposition 4.3 (Optimal sub-Weibull tail coefficient and moment condition)**.** *Let $\theta > 0$ and let X be a random variable satisfying the following asymptotic equivalence on moments*[3]

$$\|X\|_k \asymp k^\theta.$$

Then X is sub-Weibull distributed with optimal tail parameter θ, in the sense that for any $\theta' < \theta$, X is not sub-Weibull with tail parameter θ'.

Proof. Since X satisfies Condition 2 of Proposition 4.1, $X \sim \mathsf{subW}(\theta)$. Let $\theta' < \theta$. Since $\|X\|_k \asymp k^\theta$, there does not exist any constant K_2 such that $\|X\|_k \leq K_2 k^{\theta'}$, so X is not sub-Weibull with tail proxy parameter θ'. □

It is typically assumed that the random variable X has zero mean. If this is not the case, we can always center X by subtracting the mean. Let us prove that variable centering does not change the tail parameter of sub-Weibull distribution it follows.

[3]See Definition B.1 for the asymptotic equivalence $\asymp$ definition in Supplementary material.

Lemma 4.1 (Centered variables)**.** *Centering does not harm tail properties. In particular, random variables X and $(X - \mathbb{E}[X])$ belong to the same sub-Weibull family, i.e. with the same optimal tail proxy parameter.*

The proof can be found in Supplementary material.

5 Conclusion and future work

Despite the ubiquity of deep learning throughout science, medicine and engineering, the underlying theory has not kept pace with applications for deep learning in general, and for neural networks in particular. In this paper, we have extended the state of knowledge on Bayesian neural networks by providing a characterization of the marginal prior distribution of the units. We proved that they are heavier-tailed as depth increases, and interpreted this result as a sparsity-inducing mechanism at the level of the units.

Since initialization and learning dynamics are key in modern machine learning in order to properly tune deep learning algorithms, a good implementation practice requires a proper understanding of the prior distribution at play and of the regularization it incurs.

We hope that our results will open avenues for further research. Firstly, Theorem 3.1 regards the *marginal* prior distribution of the units, while a full characterization of the joint distribution of all units $\boldsymbol{U}$ remains an open question. More specifically, a precise description of the copula defined in Equation (3) would provide valuable information about the dependence between the units, and also about the precise geometrical structure of the balls induced by that penalty. Secondly, the interpretation of our result (Section 2) is concerned with the mode a posteriori of the units, which is a point estimator. One of the benefits of the Bayesian approach to neural networks lies in its ability to provide a principled approach to uncertainty quantification, so that an interpretation of our result in terms of the full posterior distribution would be very appealing. Lastly, the practical potentialities of our results are many: to delve into Bayesian deep neural networks distributional properties and better comprehend their sparsifying mechanisms will contribute to design and understand regularization strategies to avoid overfitting and improve generalization. Future work will explore these directions.

Acknowledgments

We would like to thank Stéphane Girard for fruitful discussions on Weibull-like distributions, and Cédric Févotte for pointing out the potential relationship of our heavy-tail result with sparsity-inducing priors.

References

Bibi, A., Alfadly, M., and Ghanem, B. (2018). Analytic expressions for probabilistic moments of pl-dnn with gaussian input. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, Inc.

Cho, Y. and Saul, L. K. (2009). Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, pages 342–350.

Damianou, A. and Lawrence, N. (2013). Deep Gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215.

Duvenaud, D., Rippel, O., Adams, R., and Ghahramani, Z. (2014). Avoiding pathologies in very deep networks. In *Artificial Intelligence and Statistics*, pages 202–210.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Graves, A., Mohamed, A., and Hinton, G. E. (2013). Speech recognition with deep recurrent neural networks. In *International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649.

Hron, J., Matthews, A. G. d. G., and Ghahramani, Z. (2018). Variational Bayesian dropout: pitfalls and fixes. *arXiv preprint arXiv:1807.01969*.

Kingma, D. P., Salimans, T., and Welling, M. (2015). Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*, pages 2575–2583.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems*, pages 1097–1105.

Krogh, A. and Hertz, J. A. (1991). A simple weight decay can improve generalization. In *Neural Information Processing Systems*, pages 950–957.

Lee, J., Sohl-dickstein, J., Pennington, J., Novak, R., Schoenholz, S., and Bahri, Y. (2018). Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*.

MacKay, D. J. (1992). A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472.

Matthews, A. G. d. G., Rowland, M., Hron, J., Turner, R. E., and Ghahramani, Z. (2018a). Gaussian process behaviour in wide deep neural networks. In *Proceedings of the 6th International Conference on Learning Representations*.

Matthews, A. G. d. G., Rowland, M., Hron, J., Turner, R. E., and Ghahramani, Z. (2018b). Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*.

Molchanov, D., Ashukha, A., and Vetrov, D. (2017). Variational dropout sparsifies deep neural networks. *arXiv preprint arXiv:1701.05369*.

Neal, R. M. (1992). Bayesian training of backpropagation networks by the hybrid Monte Carlo method. Technical report, Citeseer.

Neal, R. M. (1996). *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

Polson, N. G. and Sokolov, V. (2017). Deep learning: A Bayesian perspective. *Bayesian Analysis*, 12(4):1275–1304.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Rinne, H. (2008). *The Weibull distribution: a handbook*. Chapman and Hall/CRC.

Saatci, Y. and Wilson, A. G. (2017). Bayesian GAN. In *Advances in Neural Information Processing Systems*, pages 3622–3631.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., and Lanctot, M. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

A Equivalent sub-Weibull distribution properties

Proposition A.1 (Sub-Weibull distribution)**.** *Let X be a random variable. Then the following properties are equivalent; the parameters $K_i > 0$ appearing in these properties differ from each other by at most an absolute constant factor.*

1. *The tails of X satisfy*
$$\mathbb{P}(|X| \geq x) \leq 2\exp\left(-x^{1/\theta}/K_1\right) \quad \text{for all } x \geq 0.$$
2. *The moments of X satisfy*
$$\|X\|_k = \left(\mathbb{E}[|X|^k]\right)^{1/k} \leq K_2 k^\theta \quad \text{for all } k \geq 1.$$
3. *The MGF of $X^{1/\theta}$ satisfies*
$$\mathbb{E}\left[\exp\left(\lambda^{1/\theta}X^{1/\theta}\right)\right] \leq K_2 \exp(K_3^{1/\theta}\lambda^{1/\theta})$$
for all λ such that $|\lambda| \leq \frac{1}{K_3}$.
4. *The MGF of $X^{1/\theta}$ is bounded at some point, namely*
$$\mathbb{E}\left[\exp\left(X^{1/\theta}/K_4\right)\right] \leq 2.$$

Proof. **1 ⇒ 2.** Assume property **1** holds. Applying the integral identity for $|X|^k$, we obtain

$$\begin{aligned}\mathbb{E}\left[|X|^k\right] &= \int_0^\infty \mathbb{P}\left(|X|^k > x\right)dx \\ &= \int_0^\infty \mathbb{P}\left(|X| > x^{1/k}\right)dx \\ &\leq \int_0^\infty 2\exp\left(-x^{1/(k\theta)}/K_1\right)dx \\ &= 2K_1^{k\theta}k\theta\int_0^\infty e^{-u}u^{k\theta-1}du = 2K_1^{k\theta}k\theta\,\Gamma(k\theta) \\ &\sim K_1^{k\theta}k\theta\,(k\theta-1)^{k\theta-1} \sim (K_1 k\theta)^{k\theta}.\end{aligned}$$

Taking the k-th root of the expression above yields property **2**

$$\|X\|_k \lesssim (K_1\theta)^\theta k^\theta \leq K_2 k^\theta,$$

with $K_2 = (K_1\theta)^\theta$.

2 ⇒ 3. Assume property **2** holds. Recalling the Taylor series expansion of the exponential function, we obtain

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda^{1/\theta}X^{1/\theta}\right)\right] &= \mathbb{E}\left[1 + \sum_{k=1}^\infty \frac{(\lambda^{1/\theta}|X|^{1/\theta})^k}{k!}\right] \\ &= 1 + \sum_{k=1}^\infty \frac{\lambda^{k/\theta}\mathbb{E}[|X|^{k/\theta}]}{k!}.\end{aligned}$$

Property **2** guarantees that $\mathbb{E}[|X|^k] \leq K_2 k^{k/\theta}$ and $\mathbb{E}[|X|^{k/\theta}] \leq K_2(k/\theta)^k$ for some K_2. Stirling's approximation yields $k! \geq (k/e)^k$. Substituting these two bounds, we get

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda^{1/\theta}X^{1/\theta}\right)\right] &\leq \sum_{k=1}^\infty \frac{\lambda^{k/\theta}K_2(k/\theta)^k}{(k/e)^k} \\ &= \sum_{k=0}^\infty K_2(e\lambda^{1/\theta}/\theta)^k = \frac{K_2}{1 - e\lambda^{1/\theta}/\theta},\end{aligned}$$

provided that $e\lambda^{1/\theta}/\theta < 1$, in which case the geometric series above converges. To bound this quantity further, we can use the numeric inequality $\frac{1}{1-x} \leq e^{2x}$, which is valid for $x \in [0, 1/2]$. It follows that

$$\mathbb{E}\left[\exp\left(\lambda^{1/\theta}X^{1/\theta}\right)\right] \leq K_2\exp\left(2e\lambda^{1/\theta}/\theta\right)$$

for all λ satisfying $|\lambda| \leq \left(\frac{\theta}{2e}\right)^\theta$. This yields property **3** with $K_3 = (2e/\theta)^\theta$.

3 ⇒ 4. Assume property **3** holds. Take $\lambda = 1/K_4$, where $K_4 \geq K_3/(\ln 2 - \ln K_2)^\theta$. This yields property **4**.

4 ⇒ 1. Assume property **4** holds. We may assume that $K_4 = 1$. Then, by Markov's inequality and property **3**, we obtain

$$\begin{aligned}\mathbb{P}(|X| > x) &= \mathbb{P}\left(e^{|X|^{1/\theta}} > e^{x^{1/\theta}}\right) \\ &\leq \frac{\mathbb{E}[e^{|X|^{1/\theta}}]}{e^{x^{1/\theta}}} \leq 2e^{-x^{1/\theta}/K_1}.\end{aligned}$$

This proves property **1** with $K_1 = 1$. □

Remark A.1. The constant 2 that appears in some properties in Proposition A.1 does not have any special meaning. It is chosen for simplicity and can be replaced by other absolute constants.

B Intermediate lemmas

Introduce the definition of asymptotic equivalence between numeric sequences:

Definition B.1 (Asymptotic equivalence)**.** *Two sequences a_k and b_k are called asymptotic equivalent and denoted as $a_k \asymp b_k$ if there exist constants $d > 0$ and*

$D > 0$ such that

$$d \le \frac{a_k}{b_k} \le D, \quad \textit{for all } k \in \mathbb{N}. \tag{8}$$

Lemma B.1 (Gaussian moments)**.** *Let X be a normal random variable such that $X \sim \mathcal{N}(0, \sigma^2)$, then the following asymptotic equivalence holds*

$$\|X\|_k \asymp \sqrt{k}.$$

Proof. The moments of central normal absolute random variable $|X|$ are equal to

$$\begin{aligned}\mathbb{E}[|X|^k] &= \int_{\mathbb{R}} |x|^k \, p(x) \, dx \\ &= 2 \int_0^\infty x^k \, p(x) \, dx \\ &= \frac{1}{\sqrt{\pi}} \sigma^k 2^{k/2} \Gamma\left(\frac{k+1}{2}\right).\end{aligned} \tag{9}$$

We have the expression for the Gamma function

$$\Gamma(z) = \sqrt{\frac{2\pi}{z}} \left(\frac{z}{e}\right)^z \left(1 + \frac{1}{12z} + o\left(\frac{1}{z}\right)\right). \tag{10}$$

Substituting (10) into the central normal absolute moment (9), we obtain

$$\begin{aligned}\mathbb{E}[|X|^k] = \frac{1}{\sqrt{\pi}} \sigma^k 2^{k/2} \sqrt{\frac{4\pi}{k+1}} \left(\frac{k+1}{2e}\right)^{(k+1)/2} \\ \cdot \left(1 + \frac{1}{6(k+1)} + o\left(\frac{1}{k}\right)\right) \\ = \frac{2\sigma^k}{\sqrt{2e}} \left(\frac{k+1}{e}\right)^{k/2} \left(1 + \frac{1}{6(k+1)} + o\left(\frac{1}{k}\right)\right).\end{aligned}$$

Then the roots of absolute moments can be written in the form of

$$\begin{aligned}\|X\|_k &= \frac{\sigma}{e^{1/(2k)}} \sqrt{\frac{k+1}{e}} \left(1 + \frac{1}{6(k+1)} + o\left(\frac{1}{k}\right)\right)^{1/k} \\ &= \frac{\sigma}{e} \frac{\sqrt{k+1}}{e^{1/(2k)}} \left(1 + \frac{1}{6(k+1)k} + o\left(\frac{1}{k^2}\right)\right) \\ &= \frac{\sigma}{e} c_k \sqrt{k+1}.\end{aligned}$$

Here the coefficient c_k denotes

$$c_k = \frac{1}{e^{1/(2k)}} \left(1 + \frac{1}{6(k+1)k} + o\left(\frac{1}{k^2}\right)\right) \to 1,$$

with $k \to \infty$. Thus, asymptotic equivalence holds

$$\|X\|_k \asymp \sqrt{k+1} \asymp \sqrt{k}.$$

□

Lemma B.2 (Nonlinearity moments)**.** *Let X be symmetric random variable, $\phi(x)$ be nonlinear function satisfying extended envelope property, then the following asymptotic equivalence holds*

$$\|\phi(X)\|_k \asymp \|X\|_k.$$

Proof. According to extended envelope property, $\mathbb{E}[\phi(X)^k] \asymp \mathbb{E}[X^k]$. That means there exist constants d and D such that for all $k \in \mathbb{N}$ it holds

$$d \le \frac{\mathbb{E}[\phi(X)^k]}{\mathbb{E}[|X|^k]} \le D.$$

Observing that

$$d' \le d^{1/k} \le \frac{\|\phi(X)\|_k}{\|X\|_k} \le D^{1/k} \le D',$$

the bounding constants are $d' = \min\{1, d\}$, $D' = \max\{1, D\}$. It yields asymptotic equivalence

$$\|\phi(X)\|_k \asymp \|X\|_k.$$

The lemma is proved. □

Lemma B.3 (Multiplication moments)**.** *Let W and X be independent random variables such that $W \sim \mathcal{N}(0, \sigma^2)$ and for some $p > 0$ it holds*

$$\|X\|_k \asymp k^p. \tag{11}$$

Let W_i be independent copies of W, and X_i be copies of X, $i = 1, \dots, H$ with non-negative covariance between moments of copies

$$\mathit{Cov}\left[X_i^s, X_j^t\right] \ge 0, \quad \textit{for } i \ne j, \; s, t \in \mathbb{N}. \tag{12}$$

Then we have the following asymptotic equivalence

$$\Big\|\sum_{i=1}^H W_i X_i\Big\|_k \asymp k^{p+1/2}. \tag{13}$$

Proof. Let us proof the statement, using mathematical induction.

Base case: show that the statement is true for $H = 1$. For independent variables W and X, we have

$$\begin{aligned}\|WX\|_k = \left(\mathbb{E}[|WX|^k]\right)^{1/k} = \left(\mathbb{E}[|W|^k]\mathbb{E}[|X|^k]\right)^{1/k} \\ = \|W\|_k \|X\|_k.\end{aligned} \tag{14}$$

Since the random variable W follows Gaussian distribution, then Lemma B.1 implies

$$\|W\|_k \asymp \sqrt{k}. \tag{15}$$

Substituting assumption (11) and weight norm asymptotic equivalence (15) into (14) leads to the desired asymptotic equivalence (13) in case of $H = 1$.

Inductive step: show that if for $H = n-1$ the statement holds, then for $H = n$ it also holds.

Suppose for $H = n-1$ we have

$$\Big\| \sum_{i=1}^{n-1} W_i X_i \Big\|_k \asymp k^{p+1/2}. \tag{16}$$

Then, according to the covariance assumption (12), for $H = n$ we get

$$\Big\| \sum_{i=1}^{n} W_i X_i \Big\|_k^k = \Big\| \sum_{i=1}^{n-1} W_i X_i + W_n X_n \Big\|_k^k \tag{17}$$

$$\geq \sum_{j=0}^{k} C_k^j \Big\| \sum_{i=1}^{n-1} W_i X_i \Big\|_j^j \Big\| W_n X_n \Big\|_{k-j}^{k-j}. \tag{18}$$

Using the equivalence definition (Def. B.1), from the induction assumption (16) for all $j = 0, \ldots, k$ there exists absolute constant $d_1 > 0$ such that

$$\Big\| \sum_{i=1}^{n-1} W_i X_i \Big\|_j^j \geq \left(d_1 \, j^{p+1/2} \right)^j. \tag{19}$$

Recalling previous equivalence results in the base case, there exists constant $m_2 > 0$ such that

$$\Big\| W_n X_n \Big\|_{k-j}^{k-j} \geq \left(d_2 \, (k-j)^{p+1/2} \right)^{k-j}. \tag{20}$$

Substitute obtained bounds (19) and (20) into equation (17) with denoted $d = \min\{d_1, d_2\}$, obtain

$$\begin{aligned}\Big\| \sum_{i=1}^{n} W_i X_i \Big\|_k^k &\geq d^k \sum_{j=0}^{k} C_k^j \left[j^j \, (k-j)^{k-j} \right]^{p+1/2} \\ &= d^k \, k^{k(p+1/2)} \sum_{j=0}^{k} C_k^j \left[\left(\frac{j}{k}\right)^j \left(1 - \frac{j}{k}\right)^{k-j} \right]^{p+1/2}.\end{aligned} \tag{21}$$

Notice the lower bound of the following expression

$$\begin{aligned}&\sum_{j=0}^{k} C_k^j \left[\left(\frac{j}{k}\right)^j \left(1 - \frac{j}{k}\right)^{k-j} \right]^{p+1/2} \\ &\qquad \geq \sum_{j=0}^{k} \left[\left(\frac{j}{k}\right)^j \left(1 - \frac{j}{k}\right)^{k-j} \right]^{p+1/2} \geq 2.\end{aligned} \tag{22}$$

Substituting found lower bound (22) into (21), get

$$\Big\| \sum_{i=1}^{n} W_i X_i \Big\|_k^k \geq 2 \, d^k \, k^{k(p+1/2)} > d^k \, k^{k(p+1/2)}. \tag{23}$$

Now prove the upper bound. For random variables Y and Z the Holder's inequality holds

$$\begin{aligned}\|YZ\|_1 &= \mathbb{E}\left[|YZ|\right] \leq \left(\mathbb{E}\left[|Y|^2\right] \mathbb{E}\left[|Z|^2\right] \right)^{1/2} \\ &= \|YZ\|_2 \|YZ\|_2.\end{aligned}$$

Holder's inequality leads to the inequality for L^k norm

$$\|YX\|_k^k \leq \|Y\|_{2k}^k \|Z\|_{2k}^k. \tag{24}$$

Obtain the upper bound of $\|\sum_{i=1}^{n} W_i X_i\|_k^k$ from the norm property (24) for the random variables $Y = \left(\sum_{i=1}^{n-1} W_i X_i \right)^{k-j}$ and $Z = (W_n X_n)^j$

$$\Big\| \sum_{i=1}^{n} W_i X_i \Big\|_k^k = \Big\| \sum_{i=1}^{n-1} W_i X_i + W_n X_n \Big\|_k^k \tag{25}$$

$$\leq \sum_{j=0}^{k} C_k^j \Big\| \sum_{i=1}^{n-1} W_i X_i \Big\|_{2j}^j \Big\| W_n X_n \Big\|_{2(k-j)}^{k-j}. \tag{26}$$

From the induction assumption (16) for all $j = 0, \ldots, k$ there exists absolute constant $D_1 > 0$ such that

$$\Big\| \sum_{i=1}^{n-1} W_i X_i \Big\|_{2j}^j \leq \left(D_1 \, (2j)^{p+1/2} \right)^j. \tag{27}$$

Recalling previous equivalence results in the base case, there exists constant $D_2 > 0$ such that

$$\Big\| W_n X_n \Big\|_{2(k-j)}^{k-j} \leq \left(D_2 \, (2(k-j))^{p+1/2} \right)^{k-j}. \tag{28}$$

Substitute obtained bounds (27) and (28) into equation (25) with denoted $D = \max\{D_1, D_2\}$, obtain

$$\Big\| \sum_{i=1}^{n} W_i X_i \Big\|_k^k \leq D^k \sum_{j=0}^{k} C_k^j \Big[(2j)^j \, (2(k-j))^{k-j} \Big]^{p+1/2}.$$

Find an upper bound for $\left[\left(1 - \frac{j}{k}\right)^{k-j} \left(\frac{j}{k}\right)^j \right]^{p+1/2}$. Since expressions $\left(1 - \frac{j}{k}\right)$ and $\left(\frac{j}{k}\right)$ are less than 1, then $\left[\left(1 - \frac{j}{k}\right)^{k-j} \left(\frac{j}{k}\right)^j \right]^{p+1/2} < 1$ holds for all natural numbers $p > 0$. For the sum of binomial coefficients it holds the inequality $\sum_{j=0}^{k} C_k^j < 2^k$. So the final upper bound is

$$\Big\|\sum_{i=1}^{n} W_i X_i\Big\|_k^k \le 2^k\, D^k\, (2k)^{k(p+1/2)}. \tag{29}$$

Hence, taking the k-th root of (23) and (29), we have upper and lower bounds which imply the equivalence for $H = n$ and the truth of inductive step

$$d' k^{p+1/2} \le \Big\|\sum_{i=1}^{n} W_i X_i\Big\|_k \le D' k^{p+1/2},$$

where $d' = d$ and $D' = 2^{p+3/2} D$. Since both the base case and the inductive step have been performed, by mathematical induction the equivalence holds for all $H \in \mathbb{N}$

$$\Big\|\sum_{i=1}^{H} W_i X_i\Big\|_k \asymp k^{p+1/2}.$$

□

C Covariance theorem

Theorem C.1 (Non-negative covariance between hidden units). *Consider the deep neural network described in Section 3 with assumptions from Section 3.1. The covariance between hidden units of the same layer is non-negative. Moreover, for given ℓ-th hidden layer units $h^{(\ell)}$ and $\tilde{h}^{(\ell)}$, it holds*

$$Cov\left[\left(h^{(\ell)}\right)^s \left(\tilde{h}^{(\ell)}\right)^t\right] \ge 0, \text{ where } s, t \in \mathbb{N}.$$

For first hidden layer $\ell = 1$ there is equality for all s and t.

Proof. Recall the covariance definition for random variables X and Y

$$\mathrm{Cov}\,[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]. \tag{30}$$

For Gaussian random variables X and Y, $\mathrm{Cov}\,[X, Y] = 0$ means X and Y are independent.

The proof is based on induction with respect to the hidden layer number.

In the proof let us make notation simplifications: $\boldsymbol{w}_m^\ell = \boldsymbol{W}_m^\ell$ and $w_{mi}^\ell = W_{mi}^\ell$ for all $minH_\ell$. If the index m is omitted, then $\boldsymbol{w}^\ell$ is some the vectors $\boldsymbol{w}_m^\ell$, w_i^ℓ is i-th element of the vector $\boldsymbol{w}_m^\ell$.

1. First hidden layer. Consider the first hidden layer units $h^{(1)}$ and $\tilde{h}^{(1)}$. The covariance between units is equal to zero

$$\begin{aligned}\mathrm{Cov}\left[h^{(1)}, \tilde{h}^{(1)}\right] &= \mathrm{Cov}\left[\phi(g^{(1)}), \phi(\tilde{h}^{(1)})\right] \\ &= \mathrm{Cov}\left[\phi(\boldsymbol{w}^{(1)}\boldsymbol{x}), \phi(\tilde{\boldsymbol{w}}^{(1)}\boldsymbol{x})\right] = 0,\end{aligned}$$

since the weights $\boldsymbol{w}^{(1)}$ and $\tilde{\boldsymbol{w}}^{(1)}$ are from $\mathcal{N}(0, \sigma_w^2)$ and independent. Thus, the first hidden layer units are independent and its covariance (30) equals to 0.

Moreover, since $h^{(1)}$ and $\tilde{h}^{(1)}$ are independent, then $\left(h^{(1)}\right)^s$ and $\left(\tilde{h}^{(1)}\right)^t$ are also independent and it holds

$$\mathrm{Cov}\left[\left(h^{(1)}\right)^s \left(\tilde{h}^{(1)}\right)^t\right] = 0, \text{ where } s, t \in \mathbb{N}.$$

2. Next hidden layers. Assume that $(\ell-1)$-th hidden layer has $H_{\ell-1}$ hidden units, where $\ell > 1$. Then ℓ-th hidden layer pre-nonlinearity is equal to

$$g^{(\ell)} = \sum_{i=1}^{H_{\ell-1}} w_i^{(\ell)} h_i^{(\ell-1)}. \tag{31}$$

We want to prove that the covariance (30) between ℓ-th hidden layer pre-nonlinearities is non-negative. Let us show firstly the idea of the proof in the case $H_{\ell-1} = 1$ and then briefly the proof for any finite $H_{\ell-1} > 1$, $H_{\ell-1} \in \mathbb{N}$.

2.1 One hidden unit. In the case $H_{\ell-1} = 1$, the covariance (30) can be written in the form of

$$\mathrm{Cov}\left[\left(w^{(\ell)} h^{(\ell-1)}\right)^s, \left(\tilde{w}^{(\ell)} h^{(\ell-1)}\right)^t\right] \tag{32}$$

$$\begin{aligned}&= \mathbb{E}\left[\left(w^{(\ell)}\right)^s\right]\mathbb{E}\left[\left(\tilde{w}^{(\ell)}\right)^t\right]\mathbb{E}\left[\left(h^{(\ell-1)}\right)^{s+t}\right] \\ &- \mathbb{E}\left[\left(w^{(\ell)}\right)^s\right]\mathbb{E}\left[\left(\tilde{w}^{(\ell)}\right)^t\right]\mathbb{E}\left[\left(h^{(\ell-1)}\right)^s\right]\mathbb{E}\left[\left(h^{(\ell-1)}\right)^t\right]\end{aligned} \tag{33}$$

Regarding that the weighs are zero-mean distributed, its moments are equal to zero with odd order. When s (or t) is odd, the weight moment equals to zero $\mathbb{E}\left[\left(w^{(\ell)}\right)^s\right] = 0$ (or of order t) and both terms in equation (32) are equal to zero. Consider even moment orders $s = 2s_1$ and $t = 2t_1$, where $s_1, t_1 \in \mathbb{N}$.

For obtaining the non-negative covariance, we need to prove that

$$\mathbb{E}\left[\left(h^{(\ell-1)}\right)^{2(s_1+t_1)}\right] \ge \mathbb{E}\left[\left(h^{(\ell-1)}\right)^{2s_1}\right]\mathbb{E}\left[\left(h^{(\ell-1)}\right)^{2t_1}\right].$$

Since a function $f(x_1, x_2) = x_1 x_2$ is convex for $x_1 \ge 0$

and $x_2 \geq 0$, then, taking $x_1 = \left(h^{(\ell-1)}\right)^{2s_1}$ and $x_2 = \left(h^{(\ell-1)}\right)^{2t_1}$, we have

$$\mathbb{E}[f(x_1, x_2)] = \mathbb{E}\left[\left(h^{(\ell-1)}\right)^{2(t_1+s_1)}\right] \tag{34}$$

and

$$f(\mathbb{E}[x_1], \mathbb{E}[x_2]) = \mathbb{E}\left[\left(h^{(\ell-1)}\right)^{2s_1}\right]\mathbb{E}\left[\left(h^{(\ell-1)}\right)^{2t_1}\right]. \tag{35}$$

According to Jensen's inequality for convex function f, the expression (34) is more or equal to the expression (35) and the condition we need (34) is satisfied.

2.1. H hidden units. Now let us consider the covariance between pre-nonlinearities (31) for $H_{\ell-1} = H > 1$. Raise the sum in the brackets to the power

$$\begin{aligned}
&\Big(\sum_{i=1}^{H} w_i^{(\ell)} h_i^{(\ell-1)}\Big)^s = \\
&= \sum_{n_H=0}^{s} C_s^{n_H} \left(w_H^{(\ell)} h_H^{(\ell-1)}\right)^{n_H} \Big(\sum_{i=1}^{H-1} w_i^{(\ell)} h_i^{(\ell-1)}\Big)^{s-n_H} \\
&= \cdots = \\
&= \sum_{n_H=0}^{s} \sum_{n_{H-1}=0}^{s-n_H} \cdots \sum_{n_1=0}^{s-\sum_{i=2}^H n_i} C_{n_1:n_H} \cdot \\
&\quad \cdot \prod_{i=2}^{H}\left[\left(w_i^{(\ell)} h_i^{(\ell-1)}\right)^{n_i}\right]\left(w_1^{(\ell)} h_1^{(\ell-1)}\right)^{s-\sum_{i=1}^H n_i} \\
&= \sum_{n_H=0}^{s} \sum_{n_{H-1}=0}^{s-n_H} \cdots \sum_{n_1=0}^{s-\sum_{i=2}^H n_i} C_{n_1:n_H} A_{n_1:n_H} \\
&= \sum_{n_1:n_H} C_{n_1:n_H} A_{n_1:n_H}
\end{aligned}$$

where $C_{n_1:n_H} = C_s^{n_H} \prod_{i=1}^{H-1} C^{n_i}_{s-\sum_{r=i+1}^H n_r}$. And the same way for the second bracket

$$\begin{aligned}
&\Big(\sum_{i=1}^{H} \tilde{w}_i^{(\ell)} h_i^{(\ell-1)}\Big)^t = \\
&= \sum_{m_H=0}^{t} \sum_{m_{H-1}=0}^{t-m_H} \cdots \sum_{m_1=0}^{t-\sum_{i=2}^H m_i} C_{m_1:m_H} B_{m_1:m_H} = \\
&\qquad = \sum_{m_1:m_H} C_{m_1:m_H} B_{m_1:m_H}.
\end{aligned}$$

So the covariance in our notations can be written in the form of

$$\begin{aligned}
&\mathrm{Cov}\left[\Big(\sum_{i=1}^{H_{\ell-1}} w_i^{(\ell)} h_i^{(\ell-1)}\Big)^s, \Big(\sum_{i=1}^{H_{\ell-1}} w_i'^{(\ell)} h_i^{(\ell-1)}\Big)^t\right] = \\
&= \sum_{n_1:n_H} \sum_{m_1:m_H} C_{m_1:m_H} C_{n_1:n_H} \mathbb{E}\left[A_{n_1:n_H} B_{m_1:m_H}\right] \\
&- \sum_{n_1:n_H} \sum_{m_1:m_H} C_{m_1:m_H} C_{n_1:n_H} \mathbb{E}\left[A_{n_1:n_H}\right] \mathbb{E}\left[B_{m_1:m_H}\right].
\end{aligned}$$

For covariance being non-negative it is enough to show that the difference $\mathbb{E}\left[A_{n_1:n_H} B_{m_1:m_H}\right] - \mathbb{E}\left[A_{n_1:n_H}\right]\mathbb{E}\left[B_{m_1:m_H}\right]$ is non-negative for all the numbers n_i and m_i. Consider some term with numbers $n_1, \ldots, n_H, m_1, \ldots, m_H$. Since the weights are Gaussian and independent, we have the following equation, omitting the superscript for simplicity,

$$\begin{aligned}
&\mathbb{E}\left[A_{n_1:n_H} B_{m_1:m_H}\right] = \\
&= W_{n_1:n_H} \tilde{W}_{m_1:m_H} \cdot \mathbb{E}\left[h_1^{k-\sum_{i=1}^H (n_i+m_i)} \prod_{i=2}^{H} h_i^{n_i+m_i}\right],
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}\left[A_{n_1:n_H}\right]\mathbb{E}\left[B_{m_1:m_H}\right] = \\
&= W_{n_1:n_H} \tilde{W}_{m_1:m_H} \cdot \\
&\quad \cdot \mathbb{E}\left[h_1^{s-\sum_{i=1}^H n_i} \prod_{i=2}^{H} h_i^{n_i}\right]\mathbb{E}\left[h_1^{t-\sum_{i=1}^H m_i} \prod_{i=2}^{H} h_i^{m_i}\right],
\end{aligned}$$

where $W_{n_1:n_H} \tilde{W}_{m_1:m_H}$ is the product of weights moments

$$\begin{aligned}
&W_{n_1:n_H} \tilde{W}_{m_1:m_H} = \\
&= \mathbb{E}\left[w_1^{s-\sum_{i=1}^H n_i}\right]\mathbb{E}\left[\tilde{w}_1^{t-\sum_{i=1}^H m_i}\right] \prod_{i=2}^{H} \mathbb{E}\left[w_i^{n_i}\right]\mathbb{E}\left[\tilde{w}_i^{m_i}\right].
\end{aligned}$$

For $W_{n_1:n_H} \tilde{W}_{m_1:m_H}$ not equal to zero, all the powers must be even: $2s_1 = s - \sum_{i=1}^H n_i$, $2s_2 = n_2, \ldots, 2s_H = n_H$, $2t_1 = t - \sum_{i=1}^H m_i$, $2t_2 = m_2, \ldots, 2t_H = m_H$. Now we need to prove

$$\mathbb{E}\left[\prod_{i=1}^{H} h_i^{2(s_i+t_i)}\right] \geq \mathbb{E}\left[\prod_{i=1}^{H} h_i^{2s_i}\right]\mathbb{E}\left[\prod_{i=1}^{H} h_i^{2t_i}\right] \tag{36}$$

Since a function $f(x_1, x_2) = x_1 x_2$ is convex for $x_1 \geq 0$ and $x_2 \geq 0$, then, taking $x_1 = \prod_{i=1}^H h_i^{2s_i}$ and $x_2 = \prod_{i=1}^H h_i^{2t_i}$, we have

$$\mathbb{E}\Big[f(x_1, x_2)\Big] = \mathbb{E}\left[\prod_{i=1}^{H} h_i^{2(s_i+t_i)}\right] \tag{37}$$

and

$$f\Big(\mathbb{E}[x_1], \mathbb{E}[x_2]\Big) = \mathbb{E}\left[\prod_{i=1}^{H} h_i^{2s_i}\right] \mathbb{E}\left[\prod_{i=1}^{H} h_i^{2t_i}\right]. \quad (38)$$

According to Jensen's inequality for convex function f, the expression (37) is more or equal to the expression (38) and the condition we need (36) is satisfied.

3. Post-nonlinearities.

Let show the proof for the ReLU nonlinearity.

The distribution of the ℓ-th hidden layer pre-nonlinearity $g^{(\ell)}$ is the sum of symmetric distributions, which are products of Gaussian variables $w^{(\ell)}$ and non-negative ReLU output, i.e. $(\ell-1)$-th hidden layer post-nonlinearity $h^{(\ell-1)}$. It leads that $g^{(\ell)}$ follows symmetric distribution and the following inequality

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} gg'\, p(g, g')\, dg\, dg' \geq$$
$$\geq \int_{-\infty}^{+\infty} g\, p(g)\, dg \cdot \int_{-\infty}^{+\infty} g'\, p(g')\, dg'$$

implies the same inequality for positive part

$$\int_{0}^{+\infty} \int_{0}^{+\infty} gg'\, p(g, g')\, dg\, dg' \geq$$
$$\geq \int_{0}^{+\infty} g\, p(g)\, dg \cdot \int_{0}^{+\infty} g'\, p(g')\, dg'.$$

Notice that the equality above is the ReLU function output as

$$\int_{-\infty}^{+\infty} \phi(x)\, p(x)\, dx = \int_{0}^{+\infty} x\, p(x)\, dx.$$

and for symmetric distribution we have

$$\int_{0}^{+\infty} x\, p(x)\, dx = \frac{1}{2}\mathbb{E}\left[|X|\right]. \quad (39)$$

That means if non-negative covariance is proven for pre-nonlinearities, for post-nonlinearities it is also non-negative. Omit the proof for the other nonlinearities with extended envelope property, since instead of precise equation (39), the asymptotic equivalence for moments will be used for positive part and for negative part — precise expectation expression which depends on certain nonlinearity. □